Association Rule Generation for Student Performance Analysis using Apriori Algorithm

D. Magdalene Delighta Angeline*

*Assistant Professor, Department of Computer Science and Engineering, Dr.G.U.Pope College of Engineering, Thoothukudi, Tamilnadu, INDIA. E-Mail: magdalenedelighta@gmail.com

***Abstract*—**The objective of the educational institution that is producing good results in their academic exams can be achieved by using the data mining techniques which can be applied to predict the performance of the students and to impart the quality of education in the educational institutions. Data mining is used to extract meaningful information and to develop relationships among variables stored in large data set. In this paper, Apriori algorithm is used which extracts the set of rules, specific to each class and analyzes the given data to classify the student based on their performance in academics. Students are classified based on their involvement in doing assignment, internal assessment tests, attendance etc., which helps to predict the performance of the student based on the pattern extracted from the educational database. This would help to identify the average and below average students and to improve their performance to provide good results. This analysis further helps matching organization‘s requirement with students profile to provide placement for the students. Also, the interestingness of a rule is measured using lift in itself and as a part in formulae. The range of values that lift may take is used to normalize lift so that it is more effective as a measure of interestingness. This standardization is extended to account for minimum support and confidence thresholds.

***Keywords*—**Apriori, Association Rules, Data mining, Knowledge Discovery, Rule Filtering

***Abbreviations*—**Cost-Sensitive Learning (CSL), Data Mining Extensions (DMX), Knowledge Discovery in Databases (KDD), Left-Hand Side (LHS), Right-Hand Side (RHS)

I. INTRODUCTION

EDUCATION is an essential element for the development of a country. Lack of knowledge in higher educational system could prevent system management to achieve quality in education. Data mining methodology can help associating this knowledge gaps in higher education system. A better student model yields better instruction, which leads to improved learning. More accurate skill diagnosis leads to better prediction of what a student knows which provides better assessment. Better assessment leads to more efficient learning overall. The main objectives of data mining in practice tend to be prediction and description [Agrawal et al., 1994]. Predicting performance involves variables like attendance, IAT marks and assignment grades etc. in the student database to predict the unknown values. Data mining is the core process of knowledge discovery in databases. It is the process of extracting of useful patterns from the large database. In order to analyze large amount of information, the area of Knowledge Discovery in Databases (KDD) provides techniques by which the interesting patterns are extracted. Therefore, KDD utilizes methods at the cross point of machine learning, statistics and database systems. Data mining is the application of efficient algorithms to detect the desired patterns contained within the given data.

Association rules mining is one of the data mining technique which is expected to be very useful in applications. Association rules are required to assure a minimum support and a minimum confidence at the same time. Association rule generation consists of two steps: First, minimum support is applied to the given set of item. Second, using minimum confidence and frequent itemsets rules are formed. Association Rules will allow to find out rules of the type: If A then B where A and B can be particular items, values, words, etc. An association rule is composed of two item sets:

1. Antecedent or Left-Hand Side (LHS)
2. Consequent or Right-Hand Side (RHS)

It describes the relationship between support, confidence and interestingness. The support and confidence are usually referred as interestingness measures of an association rule. Association rule mining is the process of finding all the association rules with the condition of minimum support and minimum confidence. Initially, the support and confidence values are computed for all the rules and it is then compared with the threshold values to prune with low value of support or confidence Association rules mining was proposed by Agarwal. Many algorithms for generating association rules were presented over time. Some of the popular known algorithms are Apriori, Eclat and FP-Growth which is used to mine frequent itemsets. The mining exploits infrequent data, and high lowest support and high lowest confidence values. Still, it always produces an enormous amount of rules.

This paper uses association rules to extract the student performance pattern with Apriori algorithm which will be helps the education institution to analyze the student performance and to improve it in order to provide good placement for the students. The teaching organization is responsible for producing better result and the placement of students in the industry for the internship program. But it is experiencing difficulty in analyzing the student's performance at the initial stage which could lead to a poor result in the institution. Hence, staff will face problems in increasing the student result. On the other hand, some student may find difficult to perform well in the examination. As a result, this study is conducted to enhance the student result by analyzing the pattern extracted from the association rules.

Similarly, this paper also studies the mining of association rules to extract the placement pattern which will be helpful for the industry in the placement. The educational institution also experiencing difficulty in matching organization‘s requirement with students profile for several reasons. This situation could lead to a mismatched between organization's requirement and students' background. Hence, students will face problems in giving good service to the company. On the other hand, companies as well could be facing difficulties in training the students and assigning them with a project. The placement must be based on certain criteria in order to best serve the organization and student. For example, student who lives in Chennai should not be sent to an organization located in Bangalore. This is to avoid problems in terms of accommodation, financial, and social. It has been decided that practicum students' should match the organization's requirement. However, due to the large number of students registered every semester, matching the organization with the students is a very tiresome process.

II. LITERATURE REVIEW

Data mining have been applied in various research works. One of the popular techniques used for mining data in KDD for pattern discovery is the association rule [Hipp et al., 2000]. According to Usama M. Fayyad & Gregory Piatetsky-Shapiro (1996), an association rule implies certain association relationships among a set of objects. It attracted a lot of attention in current data mining research due to its capability of discovering useful patterns for decision support, selective marketing, financial prediction, medical analysis and many other applications. The association rules technique works by finding all rules in a database that satisfies the determined minimum support and minimum confidence [Bing Liu et al., 1998].

An algorithm for association rule induction is the Apriori algorithm which proves to be the accepted data mining techniques in extracting association rules [Agrawal et al., 1994], implemented the Apriori algorithm to mine single-dimensional Boolean association rules from transactional databases. The rules generated by Apriori algorithm makes it easier for the user to understand and further apply the result [Ma et al., 2000]. Employed the association rule method specifically Apriori algorithm to identifying novel, unpredicted and exciting samples in hospital infection control. Another study by employed Apriori algorithm to generate the frequent item sets and designed the model for economic forecasting, presented their methods on modeling and inferring user's intention via data. Association rules are usually required to satisfy a user-specified minimum support and a user-specified minimum confidence at the same time.

In Kotsiantis et al., (2004), Naïve Bayes algorithm is used to predict the performance of the students and the overall accuracy is found to be 72.48%. The relationship between students university entrance examination results is studied using K-means clustering technique by which the success was studied [Erdogan & Timor, 2005]. The secondary school student's performance is predicted by analyzing the result with the data mining techniques like Decision Trees, Random Forest, Neural Net-works and Support Vector Machines. The obtained results reveal that it is possible to achieve a high predictive accuracy [Cortez & Silva, 2008]. In Alaa el-Halees (2009), the data mining technique is applied to discover association rules and the discovered association rules are sorted according to the lift value. Then EM clustering technique is applied from which the outliers are detected and the performance is predicted. The recommender system techniques for educational data mining are used for predicting the performance of the students. This technique mainly focuses on focus on reducing the information overload and act as information filters [Thai-Nghe et al., 2010]. In Thai-Nghe et al., (2011, 2011A), a recommender system is used to predict the performance of the student. The information of the individual students is used to fore-casting his/her own performance. The class imbalance in the data is solved using both resampling and cost-sensitive learning (CSL) using support vector machines by which the misclassification is reduced and the classification accuracy is improved [GB-Zadok et al., 2007; Thai-Nghe, 2010A].

Association rule technique of data mining is used in Magdalene Delighta Angeline & Samuel Peter James (2012) and this paper extracts useful information from a large set of data. Likewise, this technique is applied to students' data. In Magdalene Delighta Angeline & Samuel Peter James (2012), the techniques mentioned above are used for matching the organization with the students. This process is very demanding and involves a number of steps. In Magdalene Delighta Angeline & Samuel Peter James (2012), the association rule technique provides the extracted information. On the other hand, in this paper, the creation of Data Mining Extensions (DMX) queries and their application to the Association rule model result in acquiring specific information depending on what the teacher wants to know about students' behavior patterns. Here the discussed Apriori mend algorithm generates rules by 92.86%.

This study uses Apriori algorithm and this technique is applied to the students' performance in the academics and also in the placement. This technique is used to produce rules with 100% confidence.

III. Apriori Algorithm

Figure 1 gives the Apriori algorithm. The first pass of the algorithm simply counts item occurrences to determine the large 1-itemsets. A successive pass k contains two phases: Initially, the large itemsets L_{k-1} found in the (k-1)th pass are used to generate the candidate itemsets C_k. Then, the database is scanned and the support of candidates in C_k is counted. It is necessary to determine the candidates in C_k for quick counting that are contained in a given transaction t.

Table 1 – Apriori Algorithm

```
$L_1$= {frequent items};
for (k= 2; $L_{k-1}$ !=∅; k++) do begin
$C_k$= candidates generated from $L_{k-1}$
for each transaction t in database do
  The count that are enclosed in t of all candidates in
  $C_k$ is to be incremented
$L_k$ = candidates in $C_k$ with min_sup
end
return $\cup_k L_k$;
```

IV. Implementation

In an educational institution the overall performance of a student is determined by internal assessment as well as external assessment. Internal assessment is made on the bases of a student's assignment marks, class tests, lab work, attendance, previous semester grade and his/ her involvement in extra curriculum activities. While at the same time external assessment of a student based on marks scored in final exam. The proposed model helps to predict the students about poor, average and good based on class performance as well as class attendance from the generated rules.

4.1. Dataset

The data set used in this study was obtained from department of Computer Science, Dr.G.U.Pope College of Engineering in 2011-12.

4.2. Data Mining Process

The steps of data mining process are as follows:

4.2.1. Data Selection

The data have been generated by different reports of Internal Assessment Test (IAT), Assignment, and Personal Counseling. The initial data contains the details gathered from a number of 21 students with 15 listed attributes which include Register Number, Programme, Duration, Program Code, IAT grade, Gender, Address, City1, City2, Percentage, Assignment mark Assignment submission, Correct Response, Self Confidence, Parental Education, Financial Lack, Interest and Degree aspiration. The data contains various types of values either string or numeric value. The target is represented as analysis report. The analysis report was grouped according to three categories (Good, Average and Poor). The selected attributes are IAT grade, Assignment submission, Assignment Grade, Correct Response, Self Confidence, Interest and Degree aspiration. The data were then processed for generating rules.

4.2.2. Data Transformation

Transformation has been applied to attributes Correct Response, Assignment marks. The following rules are used to transform the assignment marks to string data.

- If the Assignment mark = 9 Till 10 THEN Replace Assignment Grade by A
- If the Assignment mark = 7 Till 8 THEN Replace Assignment Grade by B
- If the Assignment mark = 5 Till 6 THEN Replace Assignment Grade by C
- If the Assignment mark = 3 Till 4 THEN Replace Assignment Grade by D
- If the Assignment mark = 1 Till 2 THEN Replace Assignment Grade by E

Likewise, the following rules are used to transform the percentage to string data

- If the Percentage = 81 Till 90 THEN Replace Percentage by S
- If the Percentage = 75 Till 80 THEN Replace Percentage by A
- If the Percentage = 70 Till 74 THEN Replace Percentage by B
- If the Percentage = 65 Till 69 THEN Replace Percentage by C

The data was then ready to be mined using association rules.

4.2.3. Rule Generation

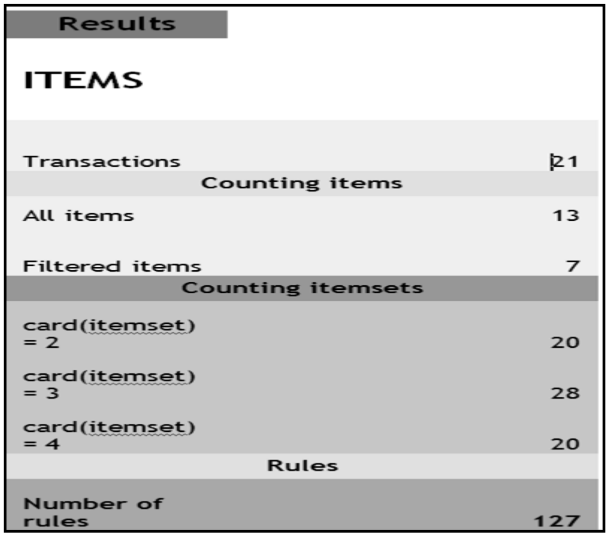
The association rules using Apriori algorithm discussed in section 3 was applied to generate rules.

V. Result Discussion

From the experiment result it is found that Apriori algorithm is used to obtain minimal rules. From the extracted pattern Apriori algorithm is found to be effective in predicting the student under three categories: good, average and poor. The numbers of transactions used in this experiment are 21. The parameters used for Apriori algorithm are minimum support, minimum confidence, and maximum rule length and lift filtering [Toscher & Jahrer, 2010]. The importance of the rule is measured using the lift value.

Table 2 – Apriori Parameters used in this System

Apriori Parameters	
Support minimum	0.33
Confidence minimum	0.75
Max rule length	4
Lift filtering	1.1

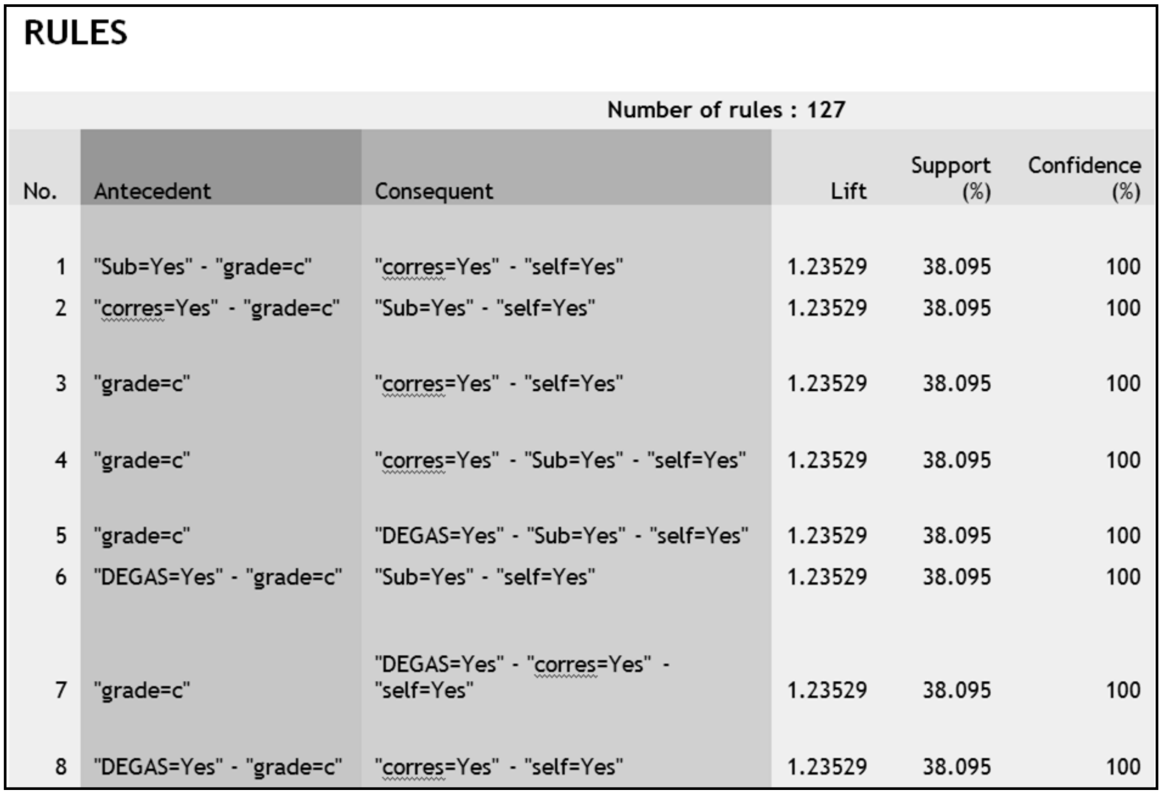


Results

ITEMS

Transactions	21
Counting items	
All items	13
Filtered items	7
Counting itemsets	
card(itemset) = 2	20
card(itemset) = 3	28
card(itemset) = 4	20
Rules	
Number of rules	127

Figure 2 – Counting Itemsets



RULES

Number of rules : 127

No.	Antecedent	Consequent	Lift	Support (%)	Confidence (%)
1	"Sub=Yes" - "grade=c"	"corres=Yes" - "self=Yes"	1.23529	38.095	100
2	"corres=Yes" - "grade=c"	"Sub=Yes" - "self=Yes"	1.23529	38.095	100
3	"grade=c"	"corres=Yes" - "self=Yes"	1.23529	38.095	100
4	"grade=c"	"corres=Yes" - "Sub=Yes" - "self=Yes"	1.23529	38.095	100
5	"grade=c"	"DEGAS=Yes" - "Sub=Yes" - "self=Yes"	1.23529	38.095	100
6	"DEGAS=Yes" - "grade=c"	"Sub=Yes" - "self=Yes"	1.23529	38.095	100
7	"grade=c"	"DEGAS=Yes" - "corres=Yes" - "self=Yes"	1.23529	38.095	100
8	"DEGAS=Yes" - "grade=c"	"corres=Yes" - "self=Yes"	1.23529	38.095	100

Figure 3 – Rules Generated using Apriori Algorithm

The output produced was evaluated in terms of accuracy. The accuracy of rules was attained according to the value of confidence value. The number of rules generated was 127 with confidence 100% and support 38.095%. Since confidence gets a value of 100 % the rule is an exact rule. The running time for the application using apriori algorithm is 15ms.The generated rules were found to be more accurate. From the generated rules the students was categorized into good, average and poor. The example of the patterns extracted from the rules is:

- If students percentage is between 80-95 THEN performance = 'Good'
- If students percentage is between 60-79 THEN performance = 'Average'
- If students percentage is below 60 THEN performance = 'Poor'

With the help of the performance report (Good, Average and Poor) of the student and the organizations criteria, the placement for the student is provided. By analysing the performance report of the student, similar training can be given. More training and concentration should be given to the poor student in order to make them pass in the examinations. Similar training and coaching should be given for the average student to perform better.

VI. Conclusion

The system has been developed to analyze the discovered rules against user's knowledge. Discovered rules can be pruned to remove redundant and insignificant rules. The scope of generated rules has been oriented to simplify the rule set and to improve the performance. The Apriori algorithm relies on downward closure property to generate all frequent itemsets that has item support above minimum support and generate confidence association rules that has confidence above the minimum confidence. The extracted rules helps to predict the performance of the students and it identify the average, below average and good students. The performance report of the student also helps to improve the result of the student. This performance enhancement will also help the entire student to get placement in various industries according to the criteria. The educational institution gets benefitted with the proposed system for their smooth and successful running of the institution. The future work can be carried out with some other data mining algorithm in terms of time factor.

References

[1] R. Agrawal, Christos Faloutsos & Arun N. Swami (1994), "Efficient Similarity Search in Sequence Databases", *Proceedings of the 4th International Conference of Foundations of Data Organization and Algorithms*, Pp. 69–84.

[2] Usama M. Fayyad & Gregory Piatetsky-Shapiro (1996), "Advances in Knowledge Discovery and Data Mining", Editors: Usama M. Fayyad & Gregory Piatetsky-Shapiro, *Cambridge, AAAI/MIT press*, Pp. 1-625.

[3] Bing Liu, Wynne Hsu & Yiming Ma (1998), "Integrating Classification and Association Rule Mining", *American Association for Artificial Intelligence*, Pp. 1–7.

[4] Y. Ma, Bing Liu, CK Wong, Philip S. Yu & SM Lee (2000), "Targeting the Right Students using Data Mining", *Proceedings of International Conference on Knowledge discovery and Data Mining*, Boston, USA, Pp. 457–464.

[5] J. Hipp, U. Guntzer & N. Gholamreza (2000), "Algorithm for Association Rule Mining: A General Survey and Comparison", *ACM SIGKDD Explorations Newsletter*, Vol. 2, No. 1, Pp. 58–64.

[6] S. Kotsiantis, C. Pierrakeas & P. Pintelas (2004), "Prediction of Student's Performance in Distance Learning using Machine Learning Techniques", *Applied Artificial Intelligence*, Vol. 18, No. 5, Pp. 411–426.

[7] Erdogan & Timor (2005), "A Data Mining Application in a Student Database", *Journal of Aeronautic and Space Technologies*, Vol. 2, No. 2, Pp. 53–57.

[8] GB-Zadok, A Hershkovitz, R Mintz & R Nachmias (2007), "Examining Online Learning Processes based on Log Files Analysis: A Case Study", *Research, Reflections and Innovations in Integrating Ict in Education*, Pp. 55–59.

[9] P. Cortez & A. Silva (2008), "Using Data Mining to Predict Secondary School Student Performance", *In EUROSIS*, Pp. 5–12.

[10] Alaa el-Halees (2009), "Mining Students Data to Analyze e-Learning Behavior: A Case Study", https://uqu.edu.sa/files2/tiny_mce/plugins/filemanager/files/30/papers/f158.pdf.

[11] A. Toscher & M.Jahrer (2010), "Collaborative Filtering Applied to Educational Data Mining", *16th ACM International Conference on Knowledge Discovery and Data Mining*, Pp. 1–11.

[12] N. Thai-Nghe (2010), "Recommender System for Predicting Student Performance", *Proceedings of the 1st Workshop on Recommender Systems for Technology Enhanced Learning*, Vol. 1, Pp. 2811–2819.

[13] N. Thai-Nghe (2010A), "Cost-Sensitive Learning Methods for Imbalanced Data", *Proceedings of the IEEE International Joint Conference on Neural Networks*, Pp. 1–8.

[14] N. Thai-Nghe (2011), "Factorization Techniques for Predicting Student Performance", *Educational Recommender Systems and Technologies: Practices and Challenges*, Pp. 129–153.

[15] N. Thai-Nghe (2011A), "Personalized Forecasting Student Performance", *Proceedings of the 11th IEEE International Conference on Advanced Learning Technologies*, Pp. 412–414.

[16] D. Magdalene Delighta Angeline & I. Samuel Peter James (2012), "Association Rule Generation using Apriori Mend Algorithm for Student's Placement", *International Journal of Emerging Sciences*, Vol. 2, No. 1, Pp. 78–86.

D. Magdalene Delighta Angeline is Assistant Professor in the Department of Computer Science and Engineering in Dr.G.U.Pope College of Engineering, Sawyerpuram, Tamilnadu, India. She obtained her Bachelor degree in Information Technology from Anna University, Chennai in the year 2007 and she obtained her Master degree in Computer and Information Technology in Manonmaniam Sundaranar University, Tirunelveli. She has over 5.7 years of Teaching Experience and published nine papers in national conference, five papers in International conferences and also published seven papers in various international journals. She also published three books. Her current area of research includes Image Processing, Neural Networks, and Data Mining.